Data Quality

Sampling error is a measure of the uncertainty in the estimate due to only observing a subset of observations from the population. The subset of observations must be chosen using probability sampling to enable sampling error to be estimated using statistical theory.

The published estimates are derived from sample data, and will differ from results derived from other samples or a census of the population. A sample and a census both experience errors classified as nonsampling errors, which often introduce systematic bias into the results. Bias is the difference between the estimate and the true value being estimated, averaged over all possible samples of the same design and size. These types of errors are not explicitly measured. Samples have sampling errors, but censuses do not. For a probability sample, this type of error can be explicitly measured. However, for any particular estimate, the total error from sampling and nonsampling error may considerably exceed the measured error.

Sampling variability

The sample selected is only one of the many possible samples that could have been selected with that same design and size, with each possible sample producing possibly different results. The relative standard error (RSE) is a measure of the sampling variability among all these possible estimates from all these possible samples, relative to the estimates. These are calculated using a delete-a-group jackknife replicate variance estimator. The RSEs in the tables can be used to derive the standard error (SE), which can then be used to create interval estimates with prescribed levels of confidence, called confidence intervals (CI). Note that the SE is in the same units as the estimate, while the RSE is a unit-less number.

The SE of the estimate is calculated by multiplying the RSE by its corresponding estimate. Note that the RSE is the measure of variability presented for all estimates in this publication except for the estimates of percent change. RSEs are also given as a percentage, and must be divided by 100 before being used to calculate the SE.

In general, for a probability sample, intervals defined by 1.645 standard errors above and below the sample estimate will contain the true population value about 90 percent of the time. Note that this definition of a 90 percent confidence interval only considers the effects of sampling, and not any of the other sources of error that can affect an estimate.

Sampling variability is also important in determining which year-to-year changes are statistically significant. The Census Bureau quality standard is that a 90% confidence interval for an estimate of change must not include zero to be considered statistically significant.

Examples of Calculating a Confidence Interval (CI)

N.B. The estimates used in the examples below are not from any particular table or cycle, but are provided to illustrate the concept.

a. Calculating a confidence interval for a specific estimate within a single survey year: Consider an estimate for a sector in a particular year from table 4a is $200,000 million and its companion RSE from table 4c is 5.0. The SE for that estimate would be calculated as:

$$\hat{\sigma}(\hat{X}_j) = \left(\frac{RSE(\hat{X}_j)}{100}\right) * \hat{X}_j$$
$$= \left(\frac{5.0}{100}\right) * \$200{,}000\ million = \$10{,}000\ million$$

The 90-percent confidence interval can be constructed by multiplying this SE by 1.645 to create the margin of error (MOE), and then adding and subtracting the MOE to the estimate. The value of 1.645 corresponds to using the Census Bureau standard of 90% confidence intervals. The 90-percent confidence interval for the estimate for this sector's total capital expenditures is:

$$\hat{X}_j \pm (1.645 * \hat{\sigma}(\hat{X}_j))$$
$$= \$200{,}000\ million\ \pm (1.645 * \$10{,}000\ million) = \$200{,}000 \pm \$16{,}450\ million$$
$$= (\$200{,}000 - \$16{,}450\ million) to\ (\$200{,}000 + \$16{,}450\ million)$$

This results in a 90 percent confidence interval from $183,550 million to $216,450 million.

Therefore, there is 90% confidence that this interval contains the true value for capital expenditures in this sector by enterprises with paid employees in the reference year.

b. Calculating a confidence interval for a percent change of an estimate between two survey years:

A confidence interval for the percent change of an estimate between two survey years can be calculated using estimates from Tables 2a and SEs from table 2b. The 90-percent confidence interval can be constructed by multiplying the SE of the percent change by 1.645 to create the Margin of Error (MOE), and then adding and subtracting the MOE to the estimate. For example, from Table 2a, the a sector's total capital expenditures estimated percent change from one year to the next is a positive 15.0 percent and from Table 2b, the standard error of this estimate is 10.0 percent.

$$15.0\% \pm (1.645 * 10.0\%) = 15.0\% \pm 16.45\%$$
$$= (15.0\% - 16.45\%)\ to\ (15.0\% + 16.45\%)$$

This results in a confidence interval of -1.45% to 31.45%.

By probability theory, 90-percent of all samples should produce an estimate of the percent change in this sector that contains the true unknown percent change. In this one observed sample, the estimate corresponds to a confidence interval of -1.45 percent to 31.45 percent. Since this confidence interval contains zero, there is *insufficient* evidence at the 90-percent confidence level to conclude that the estimated percent change was statistically different from zero, or that the change is positive. In other words, this sector showed a not statistically significant change for capital expenditures, even though the estimate of change is 15.0 percent. However, the interval is quite large, and had the estimate been slightly higher, or the standard error slightly smaller, the confidence interval might not have contained zero and shown a significant difference at the 90% confidence level.

Confidence intervals also do not consider any additional issues due to nonsampling errors (e.g. measurement errors or nonresponse biases). However, ACES confidence intervals are impacted by

nonresponse due to the weight adjustments, as discussed above. If this particular industry had a higher response rate, its estimate of change may have been statistically significant.

Examples of Calculating Differences and Percent Changes

Data for the current year along with revised data for the prior year are provided in this publication. Data users can calculate a difference, $\hat{d}_j$ and a percent change, $\widehat{PC}_j$ between the current year and prior year estimates using data on tables where the difference and percent change are not expressly given. These estimates, along with the corresponding confidence intervals, are calculated using the following equations.

The difference is calculated as:

$$\hat{d}_j = \left(\hat{X}_t - \hat{X}_{t-1}\right)$$

Where,

$\hat{X}_t$: Current year estimate of interest.

$\hat{X}_{t-1}$: Prior year estimate of interest.

The MOE for a 90-percent confidence interval on this difference is approximately:

$$MOE\left(\hat{d}_j\right) = 1.645 * \sqrt{\sigma^2\left(\hat{X}_t\right) + \sigma^2\left(\hat{X}_{t-1}\right)}$$

As an example, consider an estimate for the current year total expenditures for an industry from table 4a which is \$150,000 million with a companion RSE on Table 4c, of 4.0%. The revised prior year estimate for the same industry from Table 4b is \$130,000 million with an RSE, found in Table 4d, of 9.0%. The difference is estimated as:

$$\$150{,}000\ million - \$130{,}000\ million = \$20{,}000\ million$$

The MOE for the 90-percent confidence interval of the year-to-year change is estimated as follows, including translating the RSEs into variances by dividing the RSE by 100, multiplying by the estimate, and squaring:

$$= 1.645 * \sqrt{\left[\left(\left(\frac{4.0}{100}\right) * \$150{,}000\ million\right)^2 + \left(\left(\frac{9.0}{100}\right) * \$130{,}000\ million\right)^2\right]}$$

$$= 1.645 * \sqrt{\left[\left((0.040) * \$150{,}000\ million\right)^2 + \left((0.090) * \$130{,}000\ million\right)^2\right]}$$

$$= 1.645 * \sqrt{\$36{,}000{,}000 + \$136{,}890{,}000\ million^2}$$

$$= 1.645 * \sqrt{\$172{,}890{,}000\ million^2}$$

$$= 1.645 * \$13{,}149\ million$$

$= \$21{,}630\ million$

The 90-percent confidence interval for the difference between the two years is:

$\$20{,}000\ million \pm \$21{,}630\ million$

$(\$20{,}000 - \$21{,}630\ million)\ to\ (\$20{,}000 + \$21{,}630\ million)$

This results in a CI of -$1,630 million to $41,630 million.

Therefore, we are 90-percent confident that the difference between the prior year estimate and the current year estimate is between negative $1,630 million and $41,630 million. Since zero is in this interval, this is not sufficient evidence for a statistically significant change.

The percent change is calculated as 100 multiplied by the ratio of the difference divided by the prior estimate.

Continuing with the example from above,

$$\widehat{PC_j} = 100 * \left(\frac{\hat{d}_j}{\hat{X}_{t-1}}\right)$$

$$= 100 * \frac{\$20{,}000\ million}{\$130{,}000\ million}$$

$$= 15.4\%$$

The MOE for a 90-percent confidence interval on this percent change is estimated as:

$$MOE\left(\widehat{PC_j}\right) = 1.645 * \left(\frac{\hat{X}_t}{\hat{X}_{t-1}}\right) * \sqrt{\left(\frac{RSE_{\hat{X}_t}}{100}\right)^2 + \left(\frac{RSE_{\hat{X}_{t-1}}}{100}\right)^2}$$

$$= 1.645 * \frac{\$150{,}000\ million}{\$130{,}000\ million} * \sqrt{\left[\left(\frac{4.0}{100}\right)^2 + \left(\frac{9.0}{100}\right)^2\right]}$$

$$= 1.645 * (1.1538) * \sqrt{0.040^2 + 0.090^2}$$

$$= 1.645 * (1.1538) * \sqrt{0.0097}$$

$$= 1.645 * (1.1538) * (0.0985)$$

$$= 1.645 * 0.1136$$

$$= 0.1869$$

$$= 18.7\%$$

The 90-percent confidence interval for the percent change between the two years is:

$15.4\% \pm 18.7\%$

$(15.4\% - 18.7\%)\ to\ (15.4\% + 18.7\%)$

This gives a CI of -3.3% to 34.1%.

Because the 90-percent confidence interval contains zero, we *cannot* conclude that the percentage change from prior year to the current year estimates is a statistically significant increase at the 90-percent confidence level.

Nonsampling Error

All surveys and censuses are subject to nonsampling errors. Nonsampling errors can be attributed to many sources, including: inability to obtain information about all enterprises in the sample; inability or unwillingness on the part of respondents to provide correct information; difficulties in defining concepts; differences in the interpretation of questions; mistakes in recording or coding the data; and other errors of collection, response, coverage, and estimation for nonresponse.

Explicit measures of the effects of these nonsampling errors are not available. However, to minimize total nonsampling error, all reports were reviewed for reasonableness and consistency, and every effort was made to obtain accurate responses from all survey participants. Coverage errors, meaning errors from not including enterprises that are in-scope of the survey or mistakenly including those that are out-of-scope as eligible, may have a significant effect on the accuracy of estimates for this survey. The Business Register, a subset of which forms the sampling frame, may not contain all in-scope businesses or have incorrect values of payroll, that then affect how they are sampled and the impact of their responses through their sampling weights.

One type of nonsampling error is the error due to nonrespondents to the survey being different from respondents in substantial ways. This leads to nonresponse bias. Direct measurement of nonresponse bias, as for any bias, is difficult. Instead, some measures have been created that serve as general indicators of when nonresponse bias may be large enough to affect statistical inference.

The unit response rate (URR) is a quality measure defined as the percentage of all eligible companies that responded to the survey. If every eligible unit responded, the URR would be 100%. Companies thought to be eligible for the survey at the time of sampling may not be eligible, for instance, if the company went out of business before the start of the survey year. The URR treats all eligible companies, no matter how large or small, equally. For the 2019 ACES, the URR was 63.9% for the entire survey, and 69.6% for the ACE-1 (companies with employees) portion of the sample.

The following expression is used to calculate the URR:

$$URR = \frac{R}{S} * 100$$

URR: Unit Response Rate

R: total number of eligible companies that responded to the survey

S: total number of eligible companies sampled
The total quantity response rate (TQRR) is a quality measure defined as the percentage of the estimated total from reported or 'equivalent to reported' data. If every eligible unit responded, the TQRR would be 100%. Unlike the URR, TQRR does not treat all eligible companies equally. A company's impact on the estimates in the TQRR measure varies with its sampling weight and reported data. Each sampled unit has a sample weight reflecting other unselected companies in the population. Sampled companies in the same substratum have identical weights. Smaller weights indicate a sampled company that represents few other companies not included in the sample. Larger weights, which can be several thousand, reflect a sampled company whose reported data is used to represent the data of many similar companies that were not sampled.

In addition to sampling weights, the respondents' weights are also increased to account for companies that did not respond to the survey, which is the ACES method of adjusting for nonresponse. The proportion of the published estimates coming from respondent data using only their original unadjusted-for-nonresponse sampling weights in the total quantity response rate. In 2019, this value was 64.3% across the entire survey, and 85.9% for the ACE-1 (companies with employees) part of the sample.

The following expression is used to calculate the TQRR:

$$TQRR = \frac{\sum_{h=1}^{k} \sum_{i \in h} (W_h * X_{i,h})}{\hat{X}_{tot}}$$

TQRR: total quantity response rate

W_h: substratum-sampling weight of the h^{th} substratum

$X_{i,h}$: total capital expenditures value attributed to the i^{th} responding company of substratum h.

$\hat{X}_{tot}$: published estimate for total capital expenditures for all companies

A third quality measure used during processing is the coverage rate. The coverage rate measures the capital expenditures reported by companies with paid employees or the percentage of payroll in the sample accounted for by the respondents. The coverage rate for the 2019 ACES was 79.2%.